����������

1. ��

������������ KB◇Enterprise/Department/Personal/Agent◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇/◇◇/◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ `user_id` ◇◇◇◇RAG ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇

> ◇◇◇◇◇◇◇/◇◇◇
>
> - ◇◇◇◇ chunk ◇◇◇◇ **100 ◇◇◇◇◇◇◇◇ 1000 ◇**
> - **70% chunk ◇◇◇◇◇◇public◇**
> - ◇◇◇◇◇◇**◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇

        - ◇ public◇private ◇◇◇◇◇◇◇◇◇
        - public ◇◇◇◇ ACL filter
        - private ◇◇◇◇ ACL pre-filter◇◇◇=0◇
    4. **Hybrid 95% + rerank** ◇◇◇◇◇◇◇◇◇◇chunk 600 tokens◇◇◇◇◇◇◇◇◇ rerank ◇◇◇◇
    5. ◇◇◇◇◇◇◇◇◇ ◇◇◇◇◇◇◇◇/OCR/ASR ◇◇◇◇◇◇ text embedding
    6. ◇◇◇◇◇◇◇ **&** ◇◇◇◇

    - ◇◇◇◇◇◇◇◇◇doc_type/◇◇◇◇◇◇◇◇◇long/short◇◇◇◇◇◇◇◇◇◇◇◇◇◇
    - ◇◇◇◇chunk/keyframe◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇chunk-level Q◇◇◇◇◇◇◇ rerank ◇◇◇◇◇◇◇
    - ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ `doc_meta` / `chunk_meta`◇◇◇◇◇◇◇/◇◇/◇◇◇◇◇◇

4. ◇◇◇◇

4.1 ◇◇◇◇

- ◇◇◇◇ `user_id` ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇

4.4 ◇◇=0 ◇◇◇

- ◇◇◇◇ private ◇◇◇ ES ◇◇◇◇◇◇ `terms scope_id in user_scopes` ◇filter context◇
- ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ scope

5. ◇◇◇◇

5.1 ◇◇◇

- ◇◇◇ `kb_chunks_v1`
- public/private ◇◇◇◇◇◇◇◇◇◇ `scope_id` ◇◇
- ◇◇◇◇◇◇ `terms scope_id in user_scopes`

◇◇◇◇◇◇◇◇◇◇◇◇
◇◇◇◇◇◇◇◇public ◇◇ ACL filter ◇◇◇◇◇◇◇◇ Scale ◇◇◇

5.2 public/private ◇◇◇

- `kb_chunks_public_v1` ◇◇ scope_id=public_all
- `kb_chunks_private_v1` ◇scope_id!=public_all

◇◇◇◇◇◇◇◇◇

- public BM25 + public kNN◇◇ ACL filter◇
- private BM25 + private kNN◇◇◇ ACL filter◇
  ◇◇ + rerank ◇◇◇◇◇

6. ◇◇◇◇◇Ingestion◇◇◇

6.1 ◇◇◇

- ◇◇◇PDF/Office/HTML/Markdown/◇◇/◇◇/IM ◇
- ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
- chunk◇◇◇ 600 tokens◇overlap 80$_{120◇◇◇◇400}$800 tokens◇
- ◇◇ embedding◇768 ◇◇◇◇ `embedding_model/embedding_version`

6.2 ????????

- ???OCR → ?? → chunk → embedding
- ???ASR → transcript → chunk????+token???→ embedding
- ?????/ASR/frame → transcript → chunk???/????→ embedding
- ????????? `start_ms/end_ms`

6.3 ??

- ??? `content_hash` ?SHA1/xxhash?
- ???? `quality_score` ??????????????????????
- ????? `chunk_id` ?? ES `_id`

7. Mapping ◈◈◈JSON ◈◈◈

```
PUT kb_chunks_v1
{
  "settings": {
    "number_of_shards": 3,
    "number_of_replicas": 1,
    "refresh_interval": "10s",
    "analysis": {
      "analyzer": {
        "zh_smart": {
          "type": "ik_smart"
        },
        "zh_max": {
          "type": "ik_max_word"
        },
        "en_std": {
          "type": "standard"
        },
        "pinyin_analyzer": {
          "tokenizer": "standard",
          "filter": [
            "lowercase",
            "my_pinyin"
          ]
        }
      },
      "filter": {
        "my_pinyin": {
          "type": "pinyin",
          "keep_full_pinyin": true,
          "keep_first_letter": true,
          "keep_separate_first_letter": false,
          "keep_joined_full_pinyin": true,
          "keep_original": true,
          "limit_first_letter_length": 16,
          "remove_duplicated_term": true
        }
      }
    }
  },
```

```
"mappings": {
  "dynamic": "strict",
  "properties": {
    "kb_id": { "type": "keyword" },
    "kb_type": { "type": "keyword" },

    "scope_id": { "type": "keyword" },

    "doc_id": { "type": "keyword" },
    "chunk_id": { "type": "keyword" },
    "chunk_index": { "type": "integer" },

    "language": { "type": "keyword" },

    "title": {
      "properties": {
        "zh": {
          "type": "text",
          "analyzer": "zh_smart",
          "fields": {
            "max": { "type": "text", "analyzer": "zh_max" },
            "pinyin": { "type": "text", "analyzer": "pinyin_analyzer" },
            "keyword": { "type": "keyword", "ignore_above": 256 }
          }
        },
        "en": { "type": "text", "analyzer": "en_std" }
      }
    },
    "content": {
      "properties": {
        "zh": {
          "type": "text",
          "analyzer": "zh_smart",
          "fields": {
            "max": { "type": "text", "analyzer": "zh_max" },
            "pinyin": { "type": "text", "analyzer": "pinyin_analyzer" }
          }
        },
        "en": { "type": "text", "analyzer": "en_std" }
```

```
        }
      },

      "source_type": { "type": "keyword" },
      "doc_type": { "type": "keyword" },

      "created_at": { "type": "date" },
      "updated_at": { "type": "date" },

      "quality_score": { "type": "float" },
      "tags": { "type": "keyword" },
      "content_hash": { "type": "keyword" },

      "start_ms": { "type": "long" },
      "end_ms": { "type": "long" },

      "embedding_model": { "type": "keyword" },
      "embedding_version": { "type": "keyword" },

      "embedding": {
        "type": "dense_vector",
        "dims": 768,
        "index": true,
        "similarity": "cosine"
      }
    }
  }
}
```

8. ??/??/??/????

8.1 MVP

- shards?3~6??? 3??????
- replicas?1??????? 0?
- refresh_interval?10s

8.2 Scale

- public◈700 ◈◈◈shards=12◈8~16◈◈replicas=1
- private◈300 ◈◈◈shards=6◈4~8◈◈replicas=1

8.3 bulk ◈◈

- ◈◈ 5~15MB◈◈ 1000~5000 docs/◈◈◈ doc ◈◈◈◈
- ◈◈ 2~8◈◈◈◈◈◈◈◈◈◈
- ◈◈◈◈refresh=-1 → ◈◈◈◈◈◈ → ◈◈◈◈◈ segment

9. ◈◈◈◈

9.1 ◈◈◈◈

- BM25 size◈200◈150~400◈
- kNN◈k=150◈100250◈◈num_candidates=2000◈10006000◈
- ◈◈ topM◈150~200
- rerank topR◈60~120◈◈◈ 100◈
- ◈◈ topK◈10/20/40◈◈◈ 20◈

9.2 ??? BM25 DSL??????

```
POST kb_chunks_v1/_search
{
  "size": 200,
  "track_total_hits": false,
  "timeout": "250ms",
  "query": {
    "bool": {
      "filter": [
        { "terms": { "scope_id": ["public_all","dept_finance","project_abc"] } }
      ],
      "must": [
        {
          "multi_match": {
            "query": "?????????",
            "fields": ["title.zh^3","content.zh","title.en^2","content.en"],
            "type": "best_fields"
          }
        }
      ]
    }
  }
}
```

9.3 ◇◇◇ kNN DSL◇◇◇ pre-filter◇◇◇◇◇◇

```
POST kb_chunks_v1/_search
{
  "size": 150,
  "track_total_hits": false,
  "timeout": "400ms",
  "knn": {
    "field": "embedding",
    "query_vector": [0.01, 0.02, "..."],
    "k": 150,
    "num_candidates": 2000,
    "filter": {
      "bool": {
        "filter": [
          { "terms": { "scope_id": ["public_all","dept_finance","project_abc"] } }
        ]
      }
    }
  }
}
```

9.4 Hybrid ◇◇◇RRF◇

- ◇ BM25 ◇ kNN ◇◇◇◇◇◇◇◇◇ chunk_id◇
- RRF◇ `score = Σ 1/(k0 + rank_i)` ◇k0=60
- ◇◇◇◇ topM=200 → ◇◇ rerank topR=100

10. ◇◇◇◇◇

10.1 topK ◇◇◇◇

- topK=10◇◇◇◇◇/◇◇◇◇
- topK=20◇◇◇◇◇◇◇◇◇
- topK=40◇◇◇◇◇/◇◇◇◇◇◇/◇◇◇

10.2 ◇◇◇◇◇

- ◇◇ doc ◇◇◇◇ 2~4 ◇ chunk◇◇◇◇◇◇◇◇◇
- chunk ◇◇◇◇ doc ◇◇◇◇◇◇◇◇◇

10.3 ◇◇ chunk ◇◇◇

- ◇ topK ◇◇ doc ◇◇ chunk◇chunk_index ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ LLM
- ◇◇◇ token ◇◇◇◇◇ 8k/16k◇◇◇◇◇◇◇◇◇

11. Rerank ◇◇

- ◇◇◇◇ BM25+kNN ◇◇◇◇◇◇◇◇◇◇◇◇
- ◇◇◇**LLM rerank**◇◇◇◇◇◇◇◇◇◇◇ 20~50 ◇◇◇◇◇◇◇◇◇◇◇
- ◇◇◇◇◇◇◇◇ freshness◇updated_at ◇◇◇◇◇◇◇quality_score◇◇◇◇◇◇◇asr/ocr_confidence◇◇◇◇◇/◇◇◇◇◇◇◇◇◇
- ◇◇◇◇◇rerank ◇◇◇◇◇◇ 200~800ms◇◇◇◇/◇◇◇◇◇◇◇ topR ◇◇◇ LLM rerank◇

---

12. ◇◇◇

12.1 ◇◇◇◇

flowchart LR A[1. ◇◇◇◇
doc/pdf/html/img/audio/video] --> B[2. ◇◇/◇◇
PDF/Office/HTML/OCR/ASR] B --> C[3. ◇◇/◇◇◇
◇◇◇◇/◇◇/◇◇◇] C --> D[4. Chunk◇◇
~600 tokens + overlap] D --> E[5. ◇◇◇◇
◇◇hash/◇◇◇/◇◇◇◇] E --> F[6. Embedding◇◇
768◇+◇◇] F --> G{7. scope_id ◇◇?} G -->|public_all| H[(8. ES: kb_chunks_public_write)] G -->|non-public| I[(8. ES: kb_chunks_private_write)]
◇◇◇◇◇◇◇◇

1. ◇◇◇◇◇◇◇◇◇◇◇ PDF/Office/HTML/OCR/ASR◇◇
2. ◇◇/◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇

3. Chunk ◇◇◇
4. ◇◇◇◇◇◇◇ hash◇◇◇◇◇◇◇◇◇◇◇◇◇
5. ◇◇ Embedding◇768 ◇◇◇◇ embedding_model/version◇◇
6. ◇◇ scope_id◇public_all ◇◇◇◇◇◇◇◇◇◇
7. ◇◇◇◇◇◇◇public → `kb_chunks_public_write`◇non-public → `kb_chunks_private_write`◇

12.2 ◇◇

sequenceDiagram participant U as 1. User participant S as 2. SearchService participant ACL as 3. ACL/ScopeCache participant ES as 4. Elasticsearch participant RR as 5. RerankService participant L as 6. LLM/RAG U->>S: 1) query + user_id (+topK◇◇) S->>ACL: 2) GetUserScopes(user_id) ACL-->>S: 3) scopes (◇ public_all ◇◇◇◇◇◇◇) par BM25 S->>ES: 4a) BM25 search (filter scopes) ES-->>S: bm25_hits and kNN S->>ES: 4b) kNN search (filter scopes) ES-->>S: knn_hits end S->>S: 5) ◇◇ + RRF◇◇(topM) S->>RR: 6) rerank(topR passages/chunks) RR-->>S: rerank_scores S->>S: 7) ◇◇◇◇◇ + (◇◇)◇◇chunk◇◇ S->>L: 8) topK chunks/passages + ◇◇◇◇◇ L-->>S: 9) answer + citations S-->>U: 10) answer + citations ◇◇◇◇◇◇

1. ◇◇◇◇ query◇◇◇ user_id◇◇◇ topK ◇◇◇◇
2. ◇◇◇◇ scopes◇◇◇◇◇◇◇◇ public_all◇◇
3. ◇◇◇◇◇BM25 ◇ kNN ◇◇ scopes ◇ pre-filter◇
4. ◇◇ + RRF ◇◇◇◇ topM◇
5. ◇ topR ◇ rerank◇◇◇ cross-encoder◇◇◇ LLM rerank◇◇
6. ◇◇◇◇◇◇◇◇◇◇◇◇ chunk ◇◇◇
7. ◇ topK ◇◇◇◇◇◇◇◇ LLM/RAG◇◇◇◇◇◇◇◇◇◇

◇◇◇◇◇public/private◇◇◇◇

sequenceDiagram participant U as 1. User participant S as 2. SearchService participant ACL as 3. ACL/ScopeCache participant ESP as 4a/4b. ES Public participant ESR as 4c/4d. ES Private participant RR as 6. RerankService participant L as 8. LLM/RAG U->>S: 1) query + user_id (+topK◇◇) S->>ACL: 2) GetUserScopes(user_id) ACL-->>S: 3) scopes (◇ public_all) par Public S->>ESP: 4a) Public BM25 (no ACL) ESP-->>S: bm25_pub S->>ESP: 4b) Public kNN (no ACL) ESP-->>S: knn_pub and Private S->>ESR: 4c) Private BM25 (filter scopes) ESR--

>>S: bm25_pri S->>ESR: 4d) Private kNN (filter scopes) ESR-->>S: knn_pri end S->>S: 5) ◌◌ + RRF◌◌(topM) S->>RR: 6) rerank(topR) RR-->>S: rerank_scores S->>S: 7) ◌◌◌◌◌ + (◌◌)◌◌chunk◌◌ S->>L: 8) topK + ◌◌◌◌◌ L-->>S: 9) answer + citations S-->>U: 10) answer + citations
◌◌◌◌◌◌◌◌

1. ◌◌◌◌ query + user_id◌◌◌ topK ◌◌◌◌
2. ◌◌◌◌ scopes◌◌◌◌◌◌ public_all◌◌
3. ◌◌◌◌◌◌◌4a ◌◌ BM25◌◌ ACL◌◌4b ◌◌ kNN◌◌ ACL◌◌4c ◌◌ BM25◌◌ ACL filter◌◌4d ◌◌ kNN◌◌ ACL filter◌◌
4. ◌◌◌◌◌◌◌◌ chunk_id ◌◌◌◌ RRF ◌◌◌◌ topM◌
5. ◌ topR ◌ rerank◌◌◌ cross-encoder◌◌◌ LLM rerank◌◌
6. ◌◌◌◌◌◌◌◌◌◌◌◌ chunk ◌◌◌
7. ◌ topK ◌◌◌◌◌◌◌ LLM/RAG◌
8. LLM ◌◌ answer + citations◌
9. ◌◌◌◌◌◌

> Scale ◌◌◌ ES ◌◌◌◌ public/private ◌◌◌◌◌◌◌public ◌◌ ACL filter◌private ◌ filter◌◌

14. ◌◌◌◌◌◌◌◌◌

- ◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌ OCR/ASR◌◌◌◌◌◌◌◌◌◌◌◌◌
- ◌◌
    - ◌◌◌◌◌◌◌/◌◌/◌◌/quality_score ◌◌◌ `update` ◌◌◌◌◌◌◌ `content_hash` ◌◌◌◌
    - ◌◌◌◌◌◌◌/◌◌◌◌◌◌◌◌◌ chunk◌◌ `doc_id` ◌◌◌◌◌◌ →◌◌→◌◌ embedding→◌◌◌◌◌
    - ◌◌◌◌◌◌◌◌◌◌◌◌◌/◌◌◌/◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌◌ `chunk_id` ◌◌◌◌◌◌◌◌◌
- ◌◌◌◌◌◌
    - `chunk_id` ◌◌ `_id` ◌◌◌◌◌◌◌ `content_hash` ◌◌◌◌◌◌◌
    - ◌◌◌ bulk◌ `refresh_interval` ◌◌ 10s◌◌◌◌◌◌◌◌◌ refresh◌◌
    - ◌◌◌◌◌ `doc_id` ◌◌◌◌◌◌◌◌◌◌ chunk ◌◌◌◌◌◌◌
- ◌◌◌◌◌

        - ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇/◇◇◇◇◇◇◇
        - ◇◇◇◇◇◇◇◇◇◇◇ `kb_chunks_backup` ◇◇◇◇◇◇◇ `version` ◇◇◇◇◇◇◇◇◇◇◇
    - ◇◇◇◇◇update ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇-◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ nDCG/Recall ◇◇◇

◇◇◇◇◇◇◇

sequenceDiagram participant U as 1. Producer participant V as 2. Parser/Chunker participant S as 3. IngestService participant ES as 4. Elasticsearch participant OB as 5. Observability U->>V: 1) ◇◇◇◇◇◇(doc_id/◇◇) V->>V: 2) ◇◇/◇◇/◇◇ embedding V-->>S: 3) chunk + meta + content_hash S->>ES: 4) delete_by_query(doc_id)◇◇◇◇◇ S->>ES: 5) bulk index/update (chunk_id=_id) ES-->>S: 6) ack S->>ES: 7) optional refresh (◇◇) S-->>OB: 8) ◇◇ update ◇◇/◇◇/◇◇◇
◇◇◇◇◇◇◇◇◇◇◇◇

1. Producer ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇/◇◇/◇◇◇◇◇◇◇◇◇◇/◇◇◇◇◇◇◇◇◇◇◇
2. Parser/Chunker◇◇◇→◇◇→◇◇ embedding ◇◇◇◇◇◇◇◇◇ chunk + meta + content_hash◇
3. IngestService ◇◇◇◇◇◇◇◇◇ doc_id ◇ delete_by_query◇◇◇◇ bulk update◇
4. Bulk ◇◇◇chunk_id ◇◇ `_id` ◇◇◇◇◇ refresh ◇◇◇◇◇◇◇◇◇
5. ◇◇◇◇ doc_id ◇◇ chunk ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
6. ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ nDCG/Recall ◇◇◇

15. ◇◇◇◇◇◇◇◇◇

- ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇TTL◇
- ◇◇
    - ◇◇◇◇◇◇◇ `is_deleted=true` + `deleted_at` ◇◇◇◇◇ 7~30 ◇◇◇◇◇ filter `is_deleted=false` ◇◇◇◇◇◇
    - ◇◇◇◇◇◇ `doc_id` / `chunk_id` ◇◇ `delete_by_query` ◇◇◇◇◇◇◇ + `scroll_size` ◇◇◇
    - ◇◇◇◇◇◇◇ `scope_id` ◇◇◇◇◇◇◇◇ ACL ◇◇◇◇◇

        - ◇◇/TTL◇◇◇ ILM◇hot→warm→delete◇◇◇◇◇◇◇ `delete_by_query` ◇◇◇ `updated_at < now-ttl` ◇ `expires_at < now` ◇
    - ◇◇◇◇◇
        - ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
        - ◇◇◇◇◇◇◇◇ tombstone◇ `is_deleted=true` ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
    - ◇◇◇◇◇
        - ◇◇◇◇ `doc_id` ◇◇◇◇◇◇◇◇◇◇
        - ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇

◇◇◇◇◇◇◇

sequenceDiagram participant U as 1. Operator participant S as 2. IngestService participant ES as 3. Elasticsearch participant OB as 4. Observability U->>S: 1) ◇◇◇◇/◇◇◇◇(doc_id ◇ chunk_id) alt ◇◇◇◇ S->>ES: 2a) update is_deleted=true, deleted_at else ◇◇◇◇ S->>ES: 2b) delete_by_query(doc_id/chunk_id) end ES-->>S: 3) ack S-->>OB: 4) ◇◇◇◇◇◇/◇◇/◇◇◇◇ S->>ES: 5) ◇◇◇◇◇◇◇◇◇◇◇◇◇◇

◇◇◇◇◇◇◇◇◇◇◇◇

1. Operator ◇◇◇◇/◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇

  - ◇◇◇ alias ◇◇◇ alias ◇◇◇◇
- ◇◇◇◇◇◇◇◇ `kb_chunks_public_write_vX` / `kb_chunks_private_write_vX` ◇◇◇◇ `kb_chunks_public` / `kb_chunks_private` ◇◇◇◇◇◇◇◇◇◇◇

◇◇◇◇◇◇◇

sequenceDiagram participant OP as 1. Operator/ControlPlane participant IG as 2. IngestService participant NEW as 3. NewIndex (_write_vX) participant OLD as 4. OldIndex (◇◇◇◇◇) participant AL as 5. AliasSwitch participant OB as 6. Observability OP->>IG: 1) ◇◇◇◇◇◇/◇◇ IG->>NEW: 2) ◇◇◇◇◇◇mapping/settings◇ IG->>NEW: 3) bulk ◇◇◇◇◇refresh=-1, replicas=0◇ opt ◇◇ IG->>OLD: 4a) ◇◇◇◇◇◇◇◇◇ IG->>NEW: 4b) ◇◇◇◇◇◇◇◇◇ end IG->>IG: 5) ◇◇◇◇/◇◇◇◇ IG->>AL: 6) ◇◇◇/◇◇◇◇◇◇◇◇ AL-->>IG: 7) ack IG->>NEW: 8) ◇◇ refresh_interval/replicas IG-->>OB: 9) ◇◇◇◇◇◇◇◇◇◇ IG->>OLD: 10) ◇◇◇◇◇◇◇◇◇

◇◇◇◇

1. ◇◇◇◇◇◇◇◇◇◇ shards/replicas◇refresh_interval◇◇◇◇◇◇◇ -1◇◇◇◇/◇◇◇◇◇
2. ◇◇◇◇◇◇◇◇ Mapping/Settings ◇ `*_write_vX` ◇◇◇◇◇◇ analyzer/◇◇◇◇◇
3. ◇◇◇◇◇◇◇ replica◇refresh=-1◇◇◇ 5~15MB◇◇◇ bulk◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
4. ◇◇◇◇◇◇ IngestService ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
5. ◇◇◇

- ◇◇◇◇ doc_id/chunk ◇◇scope ◇◇◇◇◇◇
- ◇◇◇◇◇◇◇◇◇◇◇BM25/kNN/Hybrid◇◇◇◇◇◇=0◇◇◇◇◇◇◇◇

6. ◇◇◇◇◇ ingest ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇/◇◇◇◇◇◇◇◇◇◇◇ refresh_interval◇replica=1◇
7. ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ 24~72 ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ force merge ◇◇◇◇◇ segment◇

◇◇◇◇◇

- ◇◇◇◇◇◇◇reject◇◇◇◇GC◇◇◇◇◇◇ nDCG/Recall ◇◇◇
- ◇◇◇◇◇◇◇/◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇

18. ?????

- ????????/????????/??????????? scope ???/?????
- ????? doc_id/KB_id/scope_id/tag/????????????????????
  ????????

???????

sequenceDiagram participant OP as 1. Operator participant IG as 2. IngestService participant P as 3. Parser/Chunker participant ES as 4. Elasticsearch participant OB as 5. Observability OP->>IG: 1) ????????(doc_id/scope/???) IG->>P: 2) ??
??/???????? P-->>IG: 3) ?? chunk+meta+embedding IG->>ES: 4)
delete_by_query(doc_id/chunk_id) ??? IG->>ES: 5) bulk ????? refresh? IG-
>>ES: 6) scope/TTL/????????? IG-->>OB: 7) ????/???/????
?????

1. ?????????? doc_id ???????????? scan??? updated_at
   ????
2. ????????Parser/Chunker ???? chunk?meta?content_hash????
   ?? embedding/?????
3. ????? doc_id ? delete_by_query??? chunk_id ??????????
   ???scope ??????? scope?
4. Bulk ?????????????????????????? refresh?
5. ??????????????????????ACL???????????????????
6. ?????????/??/????????? scope ???????ACL/
   user_scopes??

- ?????????????????????? doc_id ???+?????????
  ????????????????/? scope ???????????????

19. ??????????

19.1 ????

- BM25/kNN ???P50/P95/P99?

- ◇◇◇◇◇rerank ◇◇
- ◇◇◇◇◇Recall@K◇nDCG◇◇◇◇◇◇◇◇◇=0◇

19.2 ◇◇◇◇

- ◇◇◇◇ query ◇◇◇>= 1~2 ◇◇
- ◇◇◇◇◇ 20 QPS◇◇◇ 100 QPS◇◇◇ 150 QPS◇◇◇◇◇◇◇◇
- ◇◇◇ topK=10/20/40 ◇ rerank on/off

19.3 ◇◇◇◇

◇ ES reject/◇◇◇◇◇◇◇◇◇◇◇◇◇

1. ◇ `num_candidates`
2. ◇ `k`
3. ◇ `topR` ◇rerank ◇◇◇◇◇
4. ◇◇◇ BM25-only◇◇◇◇◇◇

20. ◇◇◇◇◇◇◇

1. ◇◇◇◇ public ◇◇◇◇◇◇

- ◇◇◇public/private ◇◇◇

2. Hybrid 95% + rerank ◇◇◇

- ◇◇◇◇ rerank topR◇◇◇ rerank◇◇◇◇◇◇◇◇◇◇◇ topR ◇ LLM rerank◇

3. ◇◇◇◇◇◇◇

- ◇◇◇◇◇◇◇◇ + rerank ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇

21. ◇◇◇

◇◇	◇◇
ACL (Access	◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ `scope_id`

◈◈	◈◈
Control List)	◈◈◈◈◈
Alias switch	◈◈◈/◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈
asr/ ocr_confidence	◈◈◈◈◈◈◈◈◈ ASR/OCR ◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈
BM25	◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈
delete_by_query	ES ◈◈◈◈◈◈◈◈◈◈ API◈◈◈◈◈ doc_id/chunk_id ◈◈◈◈◈
Dual-write	◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈
force merge	ES ◈◈◈◈◈◈◈◈ segment ◈◈◈◈◈◈/ ◈◈◈◈◈◈◈◈◈◈◈◈◈
freshness	◈◈◈◈◈◈◈◈◈ `updated_at` ◈◈◈◈◈◈◈◈
Hybrid ◈◈	◈ BM25 ◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈ RRF◈◈◈◈◈◈◈◈◈◈◈◈◈◈
ILM (Index Lifecycle Management)	◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈ hot→warm→delete ◈◈◈◈◈◈◈◈◈
kNN / HNSW	◈◈◈◈◈◈◈◈◈◈◈◈HNSW ◈◈◈◈◈◈◈ kNN ◈◈◈◈◈◈◈
LLM rerank	◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈ 20~50 ◈◈◈◈◈◈◈◈
LRU (Least Recently Used)	◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈ scope ◈◈◈◈◈◈◈◈◈◈◈
MVP (Minimum Viable Product)	◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈ 3~6 shards◈1 replica◈refresh_interval=10s◈◈◈◈◈◈◈◈◈
nDCG (Normalized Discounted	◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈◈

◇◇	◇◇
Cumulative Gain)	
num_candidates	◇◇◇◇◇◇◇◇◇◇ 2000◇◇◇ 1000~6000 ◇◇◇◇◇◇◇◇◇◇
quality_score	◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
refresh_interval	◇◇◇◇◇MVP ◇◇ 10s◇◇◇◇◇◇◇◇◇◇◇◇
replicas	◇◇◇◇◇◇ 1◇◇◇◇◇◇◇◇◇◇ 0◇
rerank	◇◇◇◇◇◇◇◇◇◇◇◇◇◇ cross-encoder◇◇◇◇◇◇ LLM◇◇◇◇◇◇◇
RRF (Reciprocal Rank Fusion)	◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
scope	◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ `scope_id` ◇◇◇◇ `public_all`◇◇◇ ACL ◇◇◇◇◇◇
scope_id	◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
Shadow index	◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
shards	◇◇◇◇MVP ◇◇ 3~6◇◇◇ 3◇◇◇◇◇◇◇◇◇
Tombstone	◇◇◇◇◇◇◇◇ `is_deleted=true`◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
topK	◇◇◇◇◇ RAG ◇◇◇◇◇◇ 10/20/40◇◇◇ 20◇◇
topM	BM25+kNN ◇◇◇◇◇◇◇◇◇◇ 150~200◇◇◇◇◇◇ rerank ◇◇◇◇
topR	rerank ◇◇◇◇◇◇◇ 60~120◇◇◇ 100◇◇◇◇◇◇◇◇◇◇◇◇◇◇
TTL (Time To Live)	◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ scope ◇◇ 1~5 ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇
◇◇=0	◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ `scope_id` ◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇